ELICITING SUBJECTIVE PROBABILITY DISTRIBUTIONS ON CONTINUOUS VARIABLES

David A. Seaver, et al

University of Southern California

Prepared for:

Office of Naval Research
Advanced Research Projects Agency

1 August 1975

001597-3-T

USC
UNIVERSITY OF SOUTHERN CALIFORNIA

social science research institute

TECHNICAL REPORT

ELICITING SUBJECTIVE PROBABILITY DISTRIBUTIONS ON CONTINUOUS VARIABLES

DAVID A. SEAVER
DETLOF V. WINTERFELDT
WARD EDWARDS

SPONSORED BY:

ADVANCED RESEARCH PROJECTS AGENCY
DEPARTMENT OF DEFENSE

MONITORED BY:

ENGINEERING PSYCHOLOGY PROGRAMS
OFFICE OF NAVAL RESEARCH
CONTRACT NO. N00014-75-C-0487, ARPA
ORDER #2105

AUGUST 1975

SSRI RESEARCH REPORT 75-8

Social Science Research Institute
University of Southern California
Los Angeles, California 90007
213-746-6955

The Social Science Research Institute of the University of Southern California was founded on July 1, 1972 to permit USC scientists to bring their scientific and technological skills to bear on social and public policy problems. Its staff members include faculty and graduate students from many of the Departments and Schools of the University.

SSRI's research activities, supported in part from University funds and in part by various sponsors range from extremely basic to relatively applied. Most SSRI projects mix both kinds of goals — that is, they contribute to fundamental knowledge in the field of a social problem, and in doing so, help to cope with that problem. Typically, SSRI programs are interdisciplinary, drawing not only on its own staff but on the talents of others within the USC community. Each continuing program is composed of several projects; these change from time to time depending on staff and sponsor interest.

At present (Spring, 1975), SSRI has four programs:

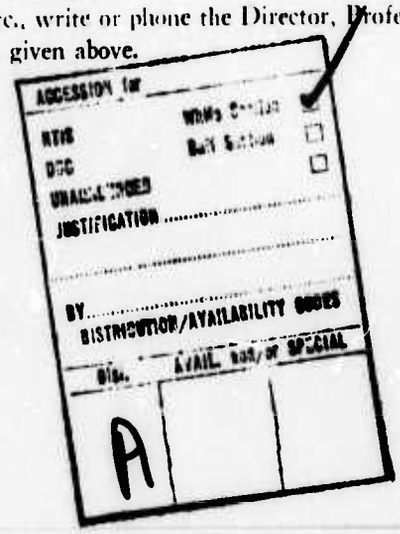
Criminal justice and juvenile delinquency. Typical projects include studies of the effect of diversion on recidivism among Los Angeles area juvenile delinquents, and evaluation of the effects of decriminalization of status offenders.

Decision analysis and social program evaluation. Typical projects include study of elicitation methods for continuous probability distributions and development of an evaluation technology for California Coastal Commission decision-making.

Program for data research. A typical project is examination of small-area crime statistics for planning and evaluation of innovations in California crime prevention programs.

Models for social phenomena. Typical projects include differential-equation models of international relations transactions and models of population flows.

SSRI anticipates continuing these four programs and adding new staff and new programs from time to time. For further information, publications, etc., write or phone the Director, Professor Ward Edwards at the address given above.



SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 001597-3-T
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): Eliciting Subjective Probability Distributions on Continuous Variables
5. TYPE OF REPORT & PERIOD COVERED: Technical
6. PERFORMING ORG. REPORT NUMBER: none

7. AUTHOR(s): David A. Seaver, Detlof v. Winterfeldt, and Ward Edwards
8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0487

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Social Science Research Institute, University of Southern California, Los Angeles, California 90007
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 197-021, ARPA Order No. 2105

11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency, 1400 Wilson Boulevard, Arlington, Virginia 22209
12. REPORT DATE: August 1, 1975
13. NUMBER OF PAGES: 23

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): Engineering Psychology Programs, Office of Naval Research, Arlington, Virginia 22217
15. SECURITY CLASS. (of this report): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for Public Release; Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Anchoring and Adjusting
Proper Scoring Rule
Fractile
Subjective Probability
Uncertainty Measures

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Five procedures for assessing subjective probability distributions over continuous variables were compared using almanac questions as stimuli. The procedures varied on the uncertainty measures used (probabilities, odds, and odds on a logarithmic scale) and the type of response required from the subjects (uncertainty measure or value of the unknown quantity). The results showed the often used fractile procedures were inferior to procedures requiring probabilities or odds as the response from subjects. The results are also discussed in terms of the "anchoring and adjustment" hypothesis.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ELICITING SUBJECTIVE PROBABILITY DISTRIBUTIONS ON CONTINUOUS VARIABLES

Technical Report
1 August 1975

David A. Seaver
Detlof v. Winterfeldt
Ward Edwards

Social Science Research Institute
University of Southern California

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Engineering Psychology Programs, Office of Naval Research under Contract No. N00014-75-C-0487, ARPA, Order #2105.

Approved for Public Release
Distribution Unlimited

SSRI Research Report 75-8

Eliciting Subjective Probability Distributions on Continuous Variables

David A. Seaver, Detlof v. Winterfeldt, and Ward Edwards

Probabilities are orderly numerical representations of personal opinions about possible events (Savage, 1954; see also Edwards, Lindman, and Savage, 1963). Such opinions must be communicated in order to be used; the process, important for many practical purposes, of requesting someone to communicate such numbers is called elicitation. Unfortunately, the truism of psychophysics that the same question, asked in two different though formally equivalent ways, will lead to different answers applies to judgments of uncertainty as it does to all other judgments.

If different elicitation procedures produce different numbers, which procedure and numbers should we believe and use? At a very abstract and philosophical level, the question is unanswerable; probabilities are judgments made by unique individuals about unique events, and so cannot be right, or wrong, or better, or worse. More practically, we can identify five properties that we should like individual probability estimates or ensembles of such estimates to have. Presumably the better estimates are on these five criteria, the more faith we will have in their validity.

1. Estimates should obey the usual laws of probability. In particular, the probabilities of an exhaustive set of mutually exclusive events should sum to 1, and probabilities of independent events should multiply.

2. Probabilities should be extreme. If elicitation method A assigns $p = .60$ to an event, while elicitation method B assigns $p = .80$, then A did

worse than B if the event later happens, and better if it does not. Murphy and Winkler (1968) call this property primary validity.

3. Probability distributions, taken over an ensemble of events, should yield relative frequencies close to the relative frequencies estimated for them. For the discrete case, for example, all events assigned probability .60 should have in common the property that about 60% of them occur. For the continuous case, a 90% credible interval over a continuous variable should have the property that about 90% of the true values of that variable fall within that interval. Murphy and Winkler call this property secondary validity. Note that in practice properties 2 and 3 can conflict. A good way of satisfying property 3 for predictions of rainfall, for example, would be to determine last year's relative frequency of rainy days and use that number as the estimated probability of rain every day this year. Obviously such a procedure, while it would do well with respect to property 3, would be very poor with respect to property 2.

4. Scores calculated from what are called proper or reproducing scoring rules (see Toda, 1963, Aczel and Pfanzagl, 1966) in effect combine properties 2 and 3. Such rules have the property that the expected value of the score is maximized if and only if the estimator correctly reports his true opinion.

5. Responsiveness to evidence should characterize good probability assessments. This criterion is difficult to state precisely. A rough statement would be that probabilities should be modified by evidence in a manner specified by Bayes's Theorem. Strictly speaking, this is simply criterion 1 restated, since Bayes's Theorem is (like virtually all other combination rules for probability) a direct consequence of the fact that probabilities sum to one and that independent probabilities multiply.

Much more vaguely interpreted, property 5 means that probability estimates should be reasonable--the meaning of that word in this context is much the same as its meaning in law.

Experimental work has been done bearing on all five criteria. Subjective probability distributions assessed by different techniques are inconsistent with each other (Schaefer and Borcherding, 1973; Stael von Holstein, 1971; Winkler, 1967). When assessed probabilities have been evaluated in terms of criterion 3 (secondary validity) (Alpert and Raiffa, 1969; Brown, 1973; Schaefer and Borcherding, 1973), the typical finding has been that they do not agree very well with the relative frequency of the actual events. Training improved the validity somewhat, but not as much as desired. Specifically, these studies found subjective probability distributions over continuous quantities to be much too tight when using fractile and equivalent assessment procedures (see Winkler, 1967, for a complete description of these procedures). That is, an unduly large percentage of the events fell into the extreme tails of the assessed distributions.

It is possible that these results are an artifact of the assessment procedures used, particularly the fractile procedure. In typical procedures subjects are asked to state a value of a random variable such that with probability p the true value will fall below that value, with probability 1-p above. Tversky and Kahnemann (1973) suggest that in judgments of this type, a cognitive process called anchoring and adjustment may occur. They hypothesize that when a subject is asked for values corresponding to specific fractiles, the subject first "anchors" on the value considered most likely, and then "adjusts" that value in the direction appropriate for the given fractile.

The adjustment process will, however, usually be insufficient, thus leading to too tight distributions. Thus for a p = .25 partition, the subject might assess the number appropriate for p = .50 and then reduce it somewhat, but not enough. Similarly, according to this argument, if the subject is given a value of the random variable and asked for the probability or odds that the true value is below the given value, the anchoring point will be 1:1 odds or probability of .50 and insufficient adjustment will lead to too flat distributions. Tversky and Kahneman present some empirical evidence that this is, in fact, what occurs.

Another possible factor contributing to the poor validity of distributions assessed with fractiles is the fact that most experimenters phrase their questions in terms of probabilities. Results from another task involving uncertainty measures as responses, probabilistic inference, have shown that responses in odds are often more valid (by criterion 5) than probability responses (Phillips and Edwards, 1966). Other results indicate that in some situations, odds on a logarithmic scale may be even more valid (see Goodman, 1973, for a review). It may be that subjects simply do not really understand the meaning of very large or very small probabilities. In addition, the cognitive adjustment process involved in the assessment task may very well depend on the measure of uncertainty used to ask or answer the questions.

This study investigates the question of how much the elicitation technique influences the validity (by criteria 3 and 4) of the assessed probability distributions. Several elicitation procedures of the fractile and the direct probability estimation type are applied, and several uncertainty measures are

used to investigate the effects of the questioning procedures and of the numerical expression of uncertainty on the validity of assessed distributions.

Method

Subjects. The Ss were 41 upper level undergraduate and graduate psychology students at California State University, Long Beach, who participated on a voluntary basis. All had some training in statistics with some exposure to the Bayesian approach.

Stimuli. Stimuli were almanac questions of the type used in the experiment by Alpert and Raiffa (1969). For example, one question was: "What was the population of Canada in 1973?" All questions involved continuous random variables. Such questions are convenient for research because the experimenter knows exact answers, while subjects have relatively vague information about them.

A questionnaire was developed for each assessment procedure. The questionnaires were self-contained; each included a complete set of instructions, examples, and the questions necessary to assess the probability distributions. Twenty distributions were assessed in each questionnaire; ten that had a percentage as the variable, e.g., the percentage of the population of California that lived in Los Angeles County, and ten that had absolute numbers as variables, e.g., the population of Canada. The reason for including two types of variables was that the percentages represented bounded variables, i.e., between 0 and 100, while the absolute numbers were only vaguely bounded.

Assessment Procedures. Five methods for assessing subjective probability distributions on continuous variables were compared. These methods varied on two dimensions: the measure of uncertainty used (odds, odds on a logarithmic scale, or probability), and the type of response required (uncertainty measures or values of the variable). A complete crossing of these variables would have yielded six experimental groups. But the use of odds on a logarithmic scale as stimulus with value of the unknown quantity as response does not seem sufficiently different from use of verbal odds as stimulus and value of the unknown quantity as response, so the former was omitted.

Elicitation methods requiring values of the unknown quantity as responses used questions of the form "What is the number of people such that your odds are 3:1 that the true population of Canada is less than that number?" (For probability groups, substitute "probability is .75" for "odds are 3:1"; for other almanac questions, change the words appropriately.) Methods requiring uncertainty measures as responses used questions of the form "What is your probability that the population of Canada is less than 130,000,000 people?" for the probability group and "Is the population of Canada more likely to be greater than or less than 130,000,000 people?" and "What are your odds?" for the odds groups. The verbal odds group simply wrote their odds in the appropriate blank while the logarithmic odds group marked their odds on a logarithmically spaced scale of odds from 1:1 to 1000:1 with a blank for odds larger than 1000:1.

This paper uses the abreviations ODDS, PROB, and LOGODDS for the methods requiring responses of odds, probabilities, and odds on a logarithmic scale respectively. Procedures requiring values of the variable as responses, the

commonly used fractile methods, are abbreviated ODDSFRAC and PROBFRAC for questions phrased in odds and probabilities respectively.

For the ODDSFRAC and PROBFRAC procedures the median, two quartiles and the .01 and .99 fractiles were assessed for each question. For the ODDS, PROB, and LOGODDS procedures, five values of the variable were given and the corresponding uncertainty measures were assessed. These five values were determined in the following manner. For the percentage variables they were randomly selected for each question from a uniform distribution between 1 and 99. For the absolute number variables, five colleagues were asked to give ranges of the variables that they were absolutely certain would contain the true value. The values given to the subjects were then selected randomly from a uniform distribution between the minimum and maximum values given by the five colleagues. These procedures were used to minimize the information given to Ss about the range of the variables. Some information was necessarily transmitted on the questions involving absolute number variables since the randomly selected values were all in some sense reasonable. However, the randomly selected values on the percentage variables did not add any information to the already known bounds.

In the ODDSFRAC and PROBFRAC procedures, Ss simply wrote in the values for the given fractiles. In the PROB procedure the responses required were probabilities that the true value was less than the given values, again simply written in an appropriate blank. The ODDS and LOGODDS procedures required that the S first state if the true value was more likely to be above or below the given value and then how much more likely, by writing down the odds in the form x:1 for the ODDS group or by marking on a logarithmic scale of odds from

1:1 to 1000:1 with a blank to fill in if the odds were larger than 1000:1 for the LOGODDS group.

The Ss were randomly assigned to the five assessment procedures, with 9, 9, 7, 8, and 8 Ss in the ODDS, PROB, LOGODDS, ODDSFRAC, and PROBFRAC groups respectively. The group sizes were unequal because some questionnaires were not returned.

Results

The basic data analyses compared actual relative frequencies with those expected from perfectly valid and unbiased distributions. Three such comparisons were made: the relative frequency of true values falling below the .01 value or above the .99 value of the cumulative distributions, called "surprises"; the relative frequency of true values falling within the interquartile ranges; and the relative frequency of true values falling below the assessed medians. For the ODDS, PROB, and LOGODDS procedures on some occasions it was not possible to determine for certain whether a true value fell within or outside the relevant range. For example, in the PROB procedure if the true value fell between values that the S had assigned probabilities of .10 and .30, it was not possible to determine whether the true value was within or outside the interquartile range. The relative frequencies for such cases were calculated in two ways; both by excluding such occurrences and by using linear interpolation on log odds to determine the location of the true value on the cumulative subjective distribution. The appropriate relative frequencies, calculated across Ss within each procedure and expressed as percentages, are presented in Tables 1, 2, and 3,

along with the number of distributions used for each calculation and 95% credible intervals on those percentages. The credible intervals were calculated

---

Insert Tables 1, 2, and 3 about here

---

using an algorithm suggested by Jackson (1974) for finding highest density regions in beta distributions and assuming a uniform prior distribution over relative frequency. These percentages can be compared with the expected percentages; 2% for Table 1 and 50% for Tables 2 and 3.

In interpreting Tables 1 and 2, it helps to remember that an excessively peaked subjective probability distribution will produce too many surprises and too few true values within the interquartile range, while an excessively flat distribution will do the opposite. The relative frequencies were calculated separately for questions with percentage variables and absolute number variables to permit easier comparison with past results that used only percentage variables. In addition, this breakdown facilitated comparisons between the fractile methods and the methods requiring uncertainty measures as responses. The results of the percentage questions are probably most closely linked to the purposes of the experiment, since no additional information was given the Ss on these questions. But the Tables show general similarity between the two kinds of results.

The results in Tables 1 and 2 can be interpreted in terms of the tightness of the assessed subjective distributions. The most striking result was the difference between the relative frequency of surprises in procedures requiring uncertainty measures as responses and procedures requiring fractiles as responses. Except for the LOGODDS procedure the former methods produced a much

Table 1

Percentages of Surprises
with 95% Credible Intervals of Those Percentages

Group	Percentage Variables				Absolute Variables				All Variables				N
	Percent Surprises		Credible Intervals		Percent Surprises		Credible Intervals		Percent Surprises		Credible Intervals		
ODDS	4.7	(5.5)	0.90	9.22	4.6	(3.3)	0.87	9.00	4.7	(4.5)	1.78	7.84	172(134)
PROB	3.3	(2.4)	0.36	7.03	5.6	(6.7)	1.43	10.43	4.5	(4.4)	1.71	7.53	179(160)
LOGODDS	18.8	(25.6)	10.07	28.12	20.9	(20.8)	11.59	30.70	19.9	(22.8)	13.34	26.61	136 (92)
ODDSFRAC	15.0		8.59	24.39	33.7		22.26	42.02	24.2		17.64	30.03	157
PROBFRAC	29.4		19.66	39.37	37.8		28.48	49.88	34.2		26.87	41.59	158

Note: Numbers in parentheses exclude questions for which interpolation was used. Columns headed "Credible Intervals" are the lower and upper bounds of the 95% credible intervals on corresponding percentages of surprises, based on all data (including interpolations).

Table 2

Percentages of True Values Falling Within
Interquartile Ranges with 95% Credible Intervals of Those Percentages

Group	Percentage Variables				Absolute Variables				All Variables				N
	Percent in Interquartile Ranges		Credible Intervals		Percent in Interquartile Ranges		Credible Intervals		Percent in Interquartile Ranges		Credible Intervals		
ODDS	48.2	(46.9)	37.71	58.78	46.0	(44.9)	35.62	56.39	47.1	(45.9)	39.67	54.53	172 (98)
PROB	63.3	(62.5)	53.40	73.10	50.6	(47.2)	40.26	60.85	57.0	(55.0)	49.75	64.17	179(109)
LOGODDS	27.5	(23.1)	17.33	38.10	34.3	(19.0)	23.24	45.67	30.9	(21.0)	23.26	38.66	136 (81)
ODDSFRAC	59.3		47.37	68.77	46.8		35.73	57.82	53.2		44.76	60.32	157
PROBFRAC	48.1		37.29	58.98	35.9		25.47	46.52	42.1		34.44	49.79	158

Note: Numbers in parentheses exclude questions for which interpolation was used. Columns headed "Credible Intervals" are the lower and upper bounds of the 95% credible intervals on corresponding percentages of true values falling within interquartile ranges, based on all data (including interpolations).

Table 3

Percentages of True Values Falling Below
Assessed Medians with 95% Credible Intervals of Those Percentages

Group	Percentage Variables				Absolute Variables				All Variables				N
	Percent Below Medians		Credible Intervals		Percent Below Medians		Credible Intervals		Percent Below Medians		Credible Intervals		
ODDS	41.2	(37.7)	30.86	51.61	47.1	(43.8)	36.74	57.55	44.2	(40.8)	36.81	51.70	172(157)
PROB	32.8	(29.6)	23.28	42.48	46.6	(46.3)	36.39	56.91	39.7	(38.0)	32.55	46.84	179(163)
LOGODDS	26.1	(20.0)	16.07	36.49	55.2	(51.8)	43.40	66.96	40.4	(35.3)	32.28	48.68	136(116)
ODDSFRAC	41.3		30.64	51.98	29.5		20.52	40.80	35.4		28.57	43.50	157
PROBFRAC	32.5		22.47	42.78	37.2		26.65	47.89	34.8		27.46	42.27	158

Note: Numbers in parentheses exclude questions for which interpolation was used. Columns headed "Credible Intervals" are the lower and upper bounds of the 95% credible intervals on the corresponding precentages of true values falling below assessed medians, based on all data (including interpolations).

smaller relative frequency of surprises, indicating flatter distributions. The difference was in the direction suggested by the anchoring and adjustment process, but the distributions assessed by the ODDS and PROB methods were not too flat; they were about right, though not quite flat enough. The use of interpolation does not seem to change the results qualitatively. If all distributions for which interpolation was required were assumed to be surprises, a quite unreasonable assumption, the relative frequency of surprises would be 24.7% and 14.5% for the ODDS and PROB procedures respectively, still at or below the surprise frequencies of the fractile procedures.

The relative frequency of true values within the interquartile range (a more stable and more important measure than surprises for most purposes) shows generally too peaked distributions except for the PROB procedure and the PROB-FRAC procedure on percentage variables. But all values are reasonably close to 50% except for the LOGODDS group, which is absurdly peaked. Table 2 shows an interaction: the use of odds in the fractile assessment procedures produced tighter distributions than the use of probabilities, while for procedures requiring uncertainty measures as responses, the converse is true. However, this conclusion would not hold up as convincingly (though it would probably be statistically significant) if the exceedingly peaked LOGODDS group were omitted from the analysis. This peculiarity of the LOGODDS group may be artifactual; the line of thought leading to that conclusion is discussed below.

The finding that too few true values fall below the assessed medians implies that the distributions as a whole are shifted along the x axis to the left of where they should be, i.e., give more probability than they should to low values and less than they should to high values.

In interpreting the credible intervals in Tables 1, 2, and 3, the assump-

tion of a uniform prior may be questioned. In many cases, e.g., the surprise frequencies, the interquartile range frequencies of the LOGODDS group, and some of the frequencies below the assessed medians, the data are striking enough so the prior is of little importance. For the other interquartile range frequencies and frequencies below the assessed medians, a more peaked prior with a mean of .50 could cause 50% to be included in credible intervals in which it is not included using a uniform prior. Thus the credible intervals with endpoints near the expected relative frequencies should be interpreted with caution.

As a further means of comparing the various assessment procedures, a proper scoring rule was applied to the assessed distributions. The scoring rule used was the continuous form of the ranked probability score (Epstein, 1969; Murphy, 1969) developed by a limiting process suggested by Brown (1970). Matheson and Winkler (1974) have illustrated the continuous ranked probability score as one of a class of scoring rules that they proved to be strictly proper. To apply the scoring rule all absolute variables were linearly transformed onto the zero to one interval to make the scores of all distributions comparable, by setting the largest value given by any $\underline{S}$ equal to one and the smallest value equal to zero unless there was a natural zero. The scoring rule then took the form

$$S = \int_0^t R^2(x)dx + \int_t^1 [1-R(x)]^2 dx$$

where t is the (transformed) true value and R(x) is the cumulative probability distribution of x. A piecewise linear approximation was used for R(x) between assessed values. There was no theoretical justification for this approximation, but because of the known insensitivity of scoring rules (von Winterfeldt and Edwards, 1973), it probably had little effect on the results.

The mean scores, presented in Table 4, were consistent with previous

---

Insert Table 4 about here

---

analyses in that the ODDS and PROB procedures had better (lower) scores. This was expected since these procedures did not produce distributions that were as much peaked as those produced by the other procedures, and also produced slightly less median displacement. It is interesting that the LOGODDS procedure, in spite of its relatively poor showing in Tables 1-3, was still preferable to either fractile procedure, according to the scoring rule. We believe this is because the scoring rule rewards probabilities that are extreme as well as close to the expected relative frequencies. (See criterion 5 above.) Apparently the extremeness of the LOGODDS procedure compensated for its poor showing compared with expected relative frequencies as evaluated by the scoring rule.

Discussion

The use of fractile methods to assess subjective probability distributions in this study led to the same excessive number of surprises found in previous studies (Alpert and Raiffa, 1969; Brown, 1973; Schaefer and Borcherding, 1973). Although training seems to improve the results, it appears other methods of assessment are needed. The ODDS and PROB procedures used in this study seem to provide more valid results. The relative frequency of true values in the tails of the distributions was much smaller for these procedures. The LOGODDS procedure produces little if any improvement over the fractiles procedures. In fact, it may be worse. In this study as in previous studies, odds assessed on a logarithmic scale seem to produce larger odds than verbal odds (see Goodman, 1973). Whether the larger odds are more valid depends on the task and task parameters. In this study they were not.

Table 4

Means and Standard Deviations from Scoring Rule

Group	Percentage Variables		Absolute Variables		All Variables	
	mean	s.d.	mean	s.d.	mean	s.d.
ODDS	.10123	.01865	.06302	.01761	.08550	.01235
PROB	.08994	.04680	.06842	.01431	.07926	.02768
LOGODDS	.10450	.02906	.08508	.01211	.09272	.01663
ODDSFRAC	.11177	.01903	.10739	.02386	.10918	.01475
PROBFRAC	.12886	.03661	.09680	.02343	.11327	.02637

Although different assessment procedures yielded distributions that differed greatly in the relative frequency of surprises, the relative frequencies of true values falling within the interquartile range did not differ substantially, except for the LOGODDS procedure. In this part of the distributions the relative frequencies were near what they should have been suggesting that in the middle range of the variables, subjective probability distributions are quite valid, independent of the assessment technique. In practical situations this is often the range of primary concern. What biases do exist may possibly be eliminated by combining the use of odds and probabilities in the assessment process, since the two measures of uncertainty seem to lead to opposite biases.

A more serious problem is the degree of the median displacement. The underestimation of both percentage and absolute number variables is not entirely consistent with previous findings. Typically low percentages have been overestimated while high percentages have been underestimated (Alpert and Raiffa, 1969; Schaefer and Borcherding, 1973), which was not the case in this study. No consistent median displacement pattern has been found on absolute number variables; Brown (1973) found overestimation and Alpert and Raiffa (1969) found underestimation. It appears that a thorough investigation of what types of questions lead to which median biases is needed.

The findings of this study seem to be generally consistent with the anchoring and adjustment process hypothesized by Tversky and Kahneman (1973). In particular, the difference in the number of surprises between the ODDS and PROB procedures and the ODDSFRAC and PROBFRAC procedures was in the direction suggested by that hypothesis. The distributions assessed by the former procedures were flatter than those assessed by the latter procedures, but ODDS and PROB distributions were not too flat, suggesting that some other process is working

in addition to the anchoring and adjustment process. Perhaps there is a real tendency to overestimate knowledge (leading to too tight distributions) in addition to the anchoring and adjustment process.

The relative frequency of true values falling within the interquartile ranges seems to tell another story. The relative tightness of these ranges showed an interaction between whether odds or probabilities were used as the measure of uncertainty and the type of response required. This suggests that if the judgments are made by anchoring and adjusting, quantitatively different adjustment processes were occurring for odds and probabilities. Apparently in the fractile procedures a larger adjustment in the value was needed to go from 1:1 odds to 3:1 odds than was needed to go from a probability of .50 to .75. Correspondingly a smaller adjustment in odds than probability was needed to adjust to some fixed value. Again it appears that the hypothesis of an anchoring and adjustment process cannot completely explain the results. Although this process does seem to play a role in the judgments required in this task, more complex processes were also occurring.

Obviously this type of sterile laboratory experiment cannot provide the ultimate answer to the question of which method of probability assessment is best for real world decision problems. What it can provide is evidence about the biases involved in various assessment procedures. The better these biases are understood, the better they can be counteracted. The practical solution will usually include a combination of various procedures incorporating many consistency checks (Spetzler and Staël von Holstein, 1972). Such processes can utilize the best aspects of each procedure while allowing probable biases to be explained and perhaps reduced or even eliminated.

References

Aczel, J. and Pfanzagl, J. Remarks on the measurement of subjective probability and information. Metrika, 1966, 2, 91-105.

Alpert, M. and Raiffa, H. A progress report on the training of probability assessors. Unpublished manuscript, Harvard University, 1969.

Brown, T.A. Probabilistic forecasts and reproducing scoring systems. The RAND Corporation, RM-6299-ARPA, June, 1970.

Brown, T.A. An experiment in probabilistic forecasting. The RAND Corporation, RM-944-ARPA, July, 1973.

Edwards, W., Lindman, H., and Savage, L. Bayesian statistical inference for psychological research. Psychological Review, 1963, 70, 193-242.

Epstein, E.S. A scoring system for probability forecasts of ranked categories. Journal of Applied Meteorology, 1969, 8, 985-987.

Goodman, B.C. Direct estimation procedures for eliciting judgments about uncertain events. Technical Report # 011313-5-T, Engineering Psychology Laboratory, University of Michigan, 1973.

Jackson, P.H. Formulae for generating highest density credibility regions. ACT Technical Bulletin No. 20, American College Testing Program, University of Iowa, February, 1974.

Matheson, J.E. and Winkler, R.L. The elicitation of continuous probability distributions. Unpublished manuscript, Stanford University, April, 1974.

Murphy, A.H. On the "ranked probability score". Journal of Applied Meteorology, 1969, 8, 988-989.

Murphy, A.H. and Winkler, R.L. Scoring rules in probability assessment and evaluation. Acta Psychologica, 1970, 34, 273-286.

Phillips, L.D. and Edwards, W. Conservatism in a simple probability inference task. Journal of Experimental Psychology, 1966, 72, 346-354.

Savage, L.J. The Foundations of Statistics. New York: Wiley, 1954.

Schaefer, R.E. and Borcherding, K. The assessment of subjective probability distributions: A training experiment. Acta Psychologica, 1973, 37, 117-129.

Spetzler, C.S. and Stael von Holstein, C.-A.S. Probability encoding in decision analysis. Paper presented at the ORSA-TIMS-AIEE 1972 Joint National Meeting, Atlantic City, N.J., 8-10 November 1972.

Stael von Holstein, C.-A.S. Two techniques for assessment of subjective probability distributions--an experimental study. Acta Psychologica. 1971, 35, 478-494.

Toda, M. Measurement of subjective probability distributions. Technical Report #3. Division of Mathematical Psychology, State College, Pennsylvania, 1963.

Tversky, A. and Kahneman, D. Judgment under uncertainty: Heuristics and Biases. Oregon Research Institute Bulletin, 1973, Volume 13, Number 1.

Winkler, R.L. The assessment of prior distributions in Bayesian analysis. Journal of the American Statistical Association, 1967, 62, 776-800.

von Winterfeldt, D. and Edwards, W. Flat maxima in linear optimization models. Technical Report #011313-4-T, Engineering Psychology Laboratory, University of Michigan, 1973.

Social Science Research Institute

Research Reports

75-1 N. Miller and G. Maruyama. Ordinal Position and Peer Popularity. January, 1975.

75-2 G. Maruyama and N. Miller. Physical Attractiveness and Classroom Acceptance. January, 1975.

75-3 J.R. Newman, S. Kirby and A.W. McEachern. Drinking Drivers and Their Traffic Records. February, 1975.

75-4 Detlof von Winterfeldt and Ward Edwards. Error in Decision Analysis: How to Create the Possibility of Large Losses by Using Dominated Strategies. April, 1975.

75-5 Peter C. Gardiner and Ward Edwards. Public Values: Multi-Attribute Utility Measurement for Social Decision Making. (Forthcoming as a chapter in *Human Judgement and Decision Processes: Formal and Mathematical Approaches*, Steven Schwartz and Martin Kaplan (eds.), summer 1975.) May, 1975.

75-6 J. Buell, H. Kagiwada, and R. Kalaba. SID: A Fortran Program for System Identification. May, 1975.

75-7 J. Robert Newman. Assesing the Reliability and Validity of Multi-Attribute Utility Procedures: An Application of the Theory of Generalizability. July, 1975.

75-8 David A. Seaver, Detlof v. Winterfeldt, and Ward Edwards. Eliciting Subjective Probability Distributions on Continuous Variables. August, 1975.